基于作答时间数据检测考生在测验过程中的速度异常 *

辛韵汐[1] 秦春影[1,2] 董圣鸿[1] 喻晓锋**[1]
（[1] 江西师范大学心理学院，南昌，330022）（[2] 南昌师范学院数学与信息科学学院，南昌，330032）

摘 要 考生在测验中出现速度异常表明考生可能存在异常的作答行为，比如提前了解了测验中部分题目的信息。由于已有方法存在对考生作答速度异常敏感性较差的缺点，研究基于作答时间数据构建了两个可以检验速度异常的统计量，将新构建的统计量在实证数据上使用，展示新方法的使用过程，并对研究结果进行了分析和讨论。基于实证数据分析的结果有针对性地设计模拟实验，与典型的符号似然比检验进行比较。研究结果表明，新构建的统计量在检验考生在考试过程中的速度差异方面有更好的表现，尤其是当考生的异常程度较低时，检验力有较大的提高。

关键词 作答时间 速度 后验概率 差异检测 贝叶斯因子

1 引言

相对于较低风险的测验，与考生利益息息相关的高风险测验出现异常考试行为的现象往往更多（van der Linden, 2009）。当前，心理测量和评估呈数字化发展趋势。对于在计算机上实施的测验，其测量过程和数据的质量控制已成为重要研究议题。高质量的测量数据是对考生的特质水平做出准确评估的前提，然而实际的测量数据中可能会存在各种各样的异常“噪音”，比如由于考生作弊等导致的异常数据（胡佳琪等 , 2020; 李亚玲 , 2021; 骆方等 , 2020; 秦春影等 , 2022; 杨志明 , 徐庆树 , 2023; Shu et al., 2013）。题目预知（item preknowledge; Sinharay, 2017a）被列为教育评估中发生的三种最广泛的考试作弊类型（题目预知，考试篡改和答案抄袭）之一（Wollack & Schoenig, 2018）。在考生出现题目预知等异常作答行为后，此时的作答数据与其正常作答时的数据在特征上往往有着显著的不同，数据中包含的异常会降低其自身及整体测验数据的质量（黄庆等 , 2024; 梁润美 , 2022; 刘玥 , 刘红云 , 2021, 2022; 童昊等 , 2022; 王丹 , 刘红云 , 2023; 王雪等 , 2024; 钟小缘等，2022; Hong et al., 2021），从而对后续的分析结果产生一系列的不良影响（Cizek & Wollack, 2017; Hong et al., 2020; Lu & Sireci, 2007; Oshima, 1994; Schnipke & Scrams, 1997）。

从已有研究来看，当考生在测验中发生题目预知等异常作答行为时，其作答数据会发生测量特征上的改变，这种改变不但体现在考生的作答得分数据，同时也会体现在作答时间数据上（Cheng & Shao, 2021; van der Linden, 2011）。作答时间数据相对于得分数据的优点主要表现在：（1）作答时间作为连续数据，有更多的统计方法可以用来分析（Cheng & Shao 2021; Wise & Kong, 2005）；（2）作答时间数据不但体现了题目的测量特征，同时也体现了考生的潜在速度信息（Marianti et al., 2014）；（3）作答时间数据的收集可以在考生没有觉察的情况下完成，对考生的影响较小（Shao, 2016）。随着计算机测验的流行，作答时间和作答得分数据一样容易被采集。作答时间数据不但可以用来对考生的潜在速度做出推断（van der Linden, 2010），还可以揭示单独使用得分信息时无法识别的有关考试特征、考生作答行为等信息（Fox et al., 2020）。除此之外，作答时间作为一个过程性数据，在测验数据分析中有着独特的优势。例如用于提高特质参数估计的准确性，促进异常作答数据的检测等（Fox & Marianti, 2016; Pan & Wollack, 2021;

* 本研究得到国家自然科学基金项目（32360208，62341207）和江西省高校人文社会科学研究项目（XL23101，XL24102）的资助。
** 通讯作者：喻晓锋，E–mail: xyu6@jxnu.edu.cn
DOI:10.16719/j.cnki.1671–6981.20260219

Sinharay & Johnson, 2020; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003）。

Wollack 和 Schoenig（2018）提到检验考生在作答不同题目集合上的能力差异是用于检测考试舞弊的六类统计方法之一。Sinharay 和 Johnson（2021）基于得分数据，对考生能力的差异进行检验，即当考生在一个题目集上估计得到的能力比其在另一个项目集上有明显差异时表明考生存在异常作答行为。已有研究中通常认为题目预知对于不同考生的影响程度相同，比如 van Krimpen-Stoop 和 Meijer（2001）及 Sinharay（2016）都认为考生在作答预知题目时会表现出“异常高能力”，将考生在作答预知题目时的能力值增加某个具体值（比如 1 或 2）来体现题目预知的影响；而 Wang 等人（2017）、Sinharay 等人（2017a, 2017b）将考生在预知题目上的正确作答概率设置为固定的 .9，从而体现考生在预知题目上的“异常高正确率”。Zhu 等人（2023）基于作答时间数据，认为考生在作答预知题目时会表现出“异常快的速度”，将考生在作答预知题目时的作答速度添加固定增量的方式来体现题目预知带来的影响。这样的假设在实际的测验情境中可能并不成立，因为题目预知只能让考生表现出题目正确作答概率或作答速度的增加，但并不能保证是相同的增加或增加到相同值。一方面，符号似然比（signal likelihood ratio, SLR）统计量在检测考生的题目预知行为时表现较好，但是存在当异常程度较低时检验力不高（Sinharay, 2017a, 2017b, 2020），对于速度差异的敏感性较差的缺点；另一方面题目预知对于不同考生的影响通常是不完全相同的，需要将题目预知对于考生的影响程度作为因素纳入考虑，尤其需要考察题目预知对考生的影响程度较低时的表现。因此，本研究拟基于作答时间数据，构建基于贝叶斯的速度差异检测方法，包括两种新的方法，分别是基于贝叶斯因子和后验概率的作答速度差异检验，考察题目预知对于考生的影响程度不同时，将它们与典型的符号似然比检验进行比较。

2 测验中的作答速度异常

不论是基于作答得分数据测量得到的考生能力，还是基于作答时间数据测量得到的考生速度，所采用的测量模型通常都有一个共同的假设，即考生在整个测验过程中的潜在特质（能力或速度）是保持固定不变的（Cizek & Wollack, 2017）。Sinharay（2021）基于作答得分数据对考生的能力差异进行检测，本质上就是对考生在作答两组题目时的能力保持固定不变这一假设进行检验。受到启发，这里拟基于作答时间数据构建检测考生题目预知行为的统计量。

在正式介绍基于作答时间差异的题目预知检测方法前，首先对新方法涉及的假设和符号进行说明。将某考生参加的测验分成两个部分，分别记为子测验 S_1 和子测验 S_2，其中考生在子测验 S_1 中不存在异常作答行为，子测验 S_2 是被怀疑存在异常作答行为的部分。用 τ_1 和 τ_2 分别表示考生在完成子测验 S_1 和 S_2 上的作答速度，对于没有题目预知的考生在这两个子测验上的作答速度差异 τ_2-τ_1 的先验分布服从均值为 0，σ'_{nc}为标准差的正态分布。而对于有题目预知的考生来说，则他在两个子测验上的作答速度差异 τ_2-τ_1 的先验分布会服从均值为μ'，σ'_c为标准差的正态分布。下面以一个描述性的实例进行说明。

图 1 描述了两类考生在作答 S_1 和 S_2 时速度差异的分布，其中实线和虚线分别表示没有 / 有速度差异 τ_2-τ_1 的密度曲线。可以看出，当考生在 S_1 和 S_2 上没有速度差异时，τ_2-τ_1 服从均值为 0 的正态分布；而这里当考生在 S_1 和 S_2 上有速度差异时，τ_2-τ_1 服从均值为 1 的正态分布，即考生在测验 S_2 的速度更快。

考生在项目上的作答得分数据反映了考生对该题目考察知识的掌握程度，体现的是考生的能力。与之不同的是，考生在题目上的作答时间 RT 能够反映考生在题目上的作答速度，考生的速度与能力之间有着互相调合的关系（van der Linden, 2007）。考生速度的差异可以通过其在题目上的作答时间数据反映出来。下面首先给出下文所涉及到的作答时间模型，然后介绍基于作答时间数据检测考生题目预知的新指标的构建过程。

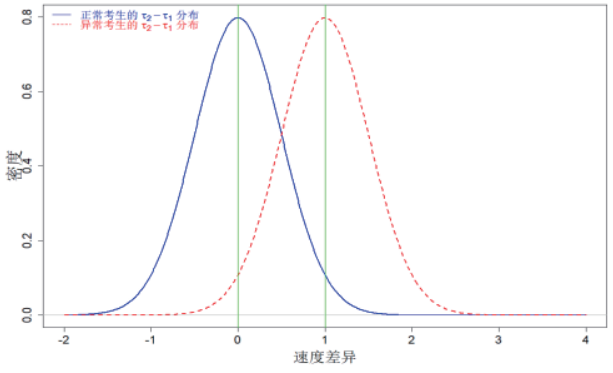

图 1 速度差异的分布

注：图中两条绿色的垂直线表示两类考生（正常考生和有题目预知的考生）在作答两个子测验上速度差异分布的均值。

3 作答时间数据的差异检验

van der Linden（2006）提出的对数正态分布作答时间模型受到广泛欢迎和使用，在该模型中作答时间呈对数正态分布，考生在考试过程中保持恒定的速度完成测验。本研究中采用两参数的对数正态时间模型，用N表示考生人数，I表示测试题目数量，则该模型下考生 n 在题目 i 上花费时间的概率密度可以表示为：

$$f(t_i;\tau_n,\alpha_i,\beta_i)=\frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_i-(\beta_i-\tau_n)\right)\right]^2\right\}, \quad (1)$$

其中，t_i 是该考生在 i 题上的作答时间；τ_n表示第 n 位考生的作答速度；β_i是题目的时间强度参数，β_i越大，考生花费在题目 i 上的时间就越多；α_i是题目的时间区分度参数，α_i越大，意味着第 i 题的作答时间分布的离散程度越小，该题在不同速度参数水平的人之间的区分性越好。Sinharay（2017a，2017b, 2020）, Sinharay 和 Johnson（2020）等研究都表明与其它已有的方法相比，符号似然比统计量在 I 类错误率和统计检测力上的表现都很优秀。因此本文只考虑将两种新的作答速度差异检验指标与符号似然比指标进行对比。

3.1 两种新的作答速度差异检验指标

3.1.1 基于作答时间数据的贝叶斯因子

一方面，基于作答得分数据，除了采用符号似然比检验检测题目预知之外，Sinharay 和 Johnson（2020）建议使用贝叶斯因子（Bayes factor; Kass & Raftery, 1995）。贝叶斯因子是一种基于贝叶斯统计的模型比较方法，它可以用来度量相对于备选模型M_1，所关注的数据是否能更好的拟合备选模型M_2时的概率。考虑本研究关注的是作答时间数据，贝叶斯因子可以表示为：

$$BF_{21}=\frac{p(\boldsymbol{t}|M_2)}{p(\boldsymbol{t}|M_1)}, \quad (2)$$

其中，$p(\boldsymbol{t}|M_2)$和$p(\boldsymbol{t}|M_1)$分别表示作答时间数据$\boldsymbol{t}$在模型M_2和M_1下的边际概率，可由下式计算得到：

$$P(\boldsymbol{t}|M_1)=\int_{\psi}p(\boldsymbol{t}|\psi,M_1)p(\psi|M_1)d\psi, \quad (3)$$

$$P(\boldsymbol{t}|M_2)=\int_{\psi}p(\boldsymbol{t}|\psi,M_2)p(\psi|M_2)d\psi, \quad (4)$$

其中，$p(\boldsymbol{t}|\psi,M_1)$是给定参数$\psi$，在模型$M_1$下的数据分布；$p(\psi|M_1)$是模型$M_1$下参数的先验分布，$P(\boldsymbol{t}|\psi,M_2)$和$p(\psi|M_2)$含义与它们类似。$BF_{21}$的值越大，支持模型$M_2$能更好拟合数据的证据就越强。有研究者提供了贝叶斯因子取值与证据强弱之间的关系（Kass & Raftery,1995）。

那么对于速度差异来说，可以将检验速度差异视为两个备选模型间的比较。模型 1 表示题目作答时间数据基于考生拥有固定的作答速度（τ）, 模型 2 中题目集$\overline{\mathbf{C}}$和$\mathbf{C}$题目作答时间数据分别基于两种不同的作答速度（τ_1 和 τ_2）。这样一来，在模型 1 和模型 2 下，考生作答时间的似然函数分别为$L(\tau;\boldsymbol{t})$和$L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)$。

因此，在检测速度差异的背景之下，贝叶斯因子可以按下式计算：

$$BF_{21}=\frac{P(\boldsymbol{t}|M_2)}{P(\boldsymbol{t}|M_1)}=\frac{\int_{\tau_1=-\infty}^{\tau_1=\infty}\int_{\tau_2=\tau_1}^{\tau_2=\infty}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2}{\int_{\tau=-\infty}^{\tau=\infty}L(\tau;\boldsymbol{t})\phi(\tau)\mathrm{d}\tau}, \quad (5)$$

其中，$p(\tau_1;\tau_2)$是 τ_1，τ_2 的联合先验分布。BF_{21}的值越大，表明数据拟合模型M_2的可能性越大。

另一方面，一些研究者（Gelman, 2014; Robert, 2007; Stern, 2005）认为相对于虚无假设，支持备择假设的直接量度指标是备择假设对应事件的后验概率，根据这个思路，下面考虑基于作答时间数据的后验概率检测考生的题目预知。

3.1.2 基于后验概率的时间差异检验

给定某考生在子测验 S_1 和 S_2 上的作答时间数据$\boldsymbol{t}_1$和$\boldsymbol{t}_2$，速度 τ_1，τ_2 的联合后验分布定义为$g(\tau_1,\tau_2|\boldsymbol{t})$。根据局部独立性假设，$g(\tau_1,\tau_2|\boldsymbol{t})$的计算公式为：

$$g(\tau_1,\tau_2|\boldsymbol{t})=\frac{L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)}{\int_{\tau_1=-\infty}^{\tau_1=\infty}\int_{\tau_2=-\infty}^{\tau_2=\infty}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2}. \quad (6)$$

根据公式 6，后验概率$P(\tau_2\geq\tau_1|\boldsymbol{t})$可计算为：

$$p(\tau_2\geq\tau_1|\boldsymbol{t})=\int_{\tau_1=-\infty}^{\tau_1}\int_{\tau_2=\tau_1}^{\infty}g(\tau_1,\tau_2|\boldsymbol{t})\mathrm{d}\tau_1\mathrm{d}\tau_2=\frac{\int_{\tau_1=-\infty}^{\tau_1}\int_{\tau_2=\tau_1}^{\tau_2}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2}{\int_{\tau_1=-\infty}^{\tau_1=\infty}\int_{\tau_2=-\infty}^{\tau_2=\infty}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2}, \quad (7)$$

可以看出，公式 7 中分子和分母的积分公式是相同的，不同的是积分区间。公式 7 中的积分需要用数值积分的方式进行计算，这里使用黎曼和近似，即上面的分子和分母可以分别按如下 8 和 9 式计算。

$$\int_{\tau_1=-\infty}^{\tau_1=\infty}\int_{\tau_2=\tau_1}^{\tau_2=\infty}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2\approx\sum_{k=1}^{K}\sum_{\substack{m=1\\ \tau_{2m}>\tau_{1k}}}^{M}L(\tau_{1k};\boldsymbol{t}_1)L(\tau_{2m};\boldsymbol{t}_2)p(\tau_{1k};\tau_{2m})\Delta_1\Delta_2, \quad (8)$$

$$\int_{\tau_1=-\infty}^{\tau_1=\infty}\int_{\tau_2=-\infty}^{\tau_2=\infty}L(\tau_1;\boldsymbol{t}_1)L(\tau_2;\boldsymbol{t}_2)p(\tau_1;\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2\approx\sum_{k=1}^{K}\sum_{m=1}^{M}L(\tau_{1k};\boldsymbol{t}_1)L(\tau_{2m};\boldsymbol{t}_2)p(\tau_{1k};\tau_{2m})\Delta_1\Delta_2, \quad (9)$$

其中，τ_{11}, ..., τ_{1k}是 K 个等间距点，τ_{21}, ..., τ_{2m}是 M 个等间距点，$\Delta_1=\tau_{1,k+1}-\tau_{1k}$, $\Delta_2=\tau_{2,m+1}-\tau_{2m}$。在本文的研究里，考虑用（-5，5）之间的 101 个等距点来近似数值积分，即步长设为 .1。基于考生的作答时间数据，公式 7 得到的概率值越大，表明考生在测验过程中完成$\overline{\mathbf{C}}$时的作答速度比完成$\mathbf{C}$时的作答

速度大的可能性越大。为了更清楚地说明三个基于作答时间数据统计量在检测题目预知时的作用，即它们在检测题目预知时对于时间差异（即速度差异）的反馈，下面通过一个描述性的模拟例子加以介绍。

3.2 一个描述性的模拟例子

为方便介绍和理解，这里我们通过一个模拟的例子对三个统计量在考生包含不同程度的异常作答时间数据时的检测结果进行分析。考虑一个包含20道题目的测验，题目的时间区分度α和时间强度β分别固定为2和1，8名考生的作答时间数据按最简单的方式设计。具体如下表1所示，即考生在测验后10题的一些题目（8位考生的预知题目数是从0到7，依次增加1）上存在预知，这样一来，每位考生存在预知的题目数量是不同的，考察检验方法在严重的题目预知和轻度的题目预知情形下的表现，可以体现检验方法的稳健性。这里采用的方法是对前10题和后10题的作答速度进行检验。考虑在测验后10题里，预知的题目数量从0到7的变化是为了让模拟情境更符合现实应用场景，因为很多时候怀疑的题目集合里只是部分题目被考生所预知。

基于表1中的作答数据，和各题目的参数，估计每位考生的3个速度参数，分别是基于前10题、后10题，和总共20题的作答时间数据，并计算每位考生的SLR统计量及其对应的p值、贝叶斯因子BF和考生的后验概率PP。对于这三个统计量，值越大表明考生在考试过程中不同子测验(题目集合）存在作答时间差异的可能性越大，即体现了该考生在测验的不同部分中存在不同的速度。

可以看出，当考生在测验的不同部分存在速度差异时，上面的三个统计量都能够给出反馈，并且随着异常程度的增大，有更大可能支持做出考生存在异常的判断。由于这个例子中的数据比较“理想”或者说是“噪音比较明显”的数据，它只是为了描述三个统计量的使用，分析的结果初步表明提出的基于贝叶斯因子和后验概率的统计量在这个数据中的分析和经典的似然比统计量同样有效。而针对符号似然比指标的局限以及三种统计指标间的差异会在更全面的模拟实验中展示。

为了评价前面三种基于作答时间数据的统计量在检测考生在测验中存在题目预知时的表现，本文首先基于一批被广泛开展研究的数据展开分析（Kasli et al., 2023; Lee, 2018; Sinharay, 2017a, 2017b, 2020），这批数据在众多关注题目预知的研究中被

表1 描述性实例中的考生作答时间

考生	题目作答时间（单位：秒）																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
2	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5
3	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5
4	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5	5
5	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5	5	5
6	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5	5	5	5
7	15	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5	5	5	5	5
8	15	15	15	15	15	15	15	15	15	15	15	15	15	5	5	5	5	5	5	5

注：红色标出的题目表示考生已经预知了该题目，导致其在该题目的作答时间变短。

表2 基于描述性实例计算的统计量值

考生	时间差异（单位：秒）	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	SLR（p）	BF	PP
1	0	-1.708	-1.708	-1.708	0（.500）	.169	.432
2	10	-1.708	-1.598	-1.653	.491（.312）	.245	.610
3	20	-1.708	-1.488	-1.598	.983（.163）	.388	.768
4	30	-1.708	-1.378	-1.543	1.474（.070）	.686	.882
5	40	-1.708	-1.269	-1.488	1.965（.025）	1.397	.949
6	50	-1.708	-1.159	-1.433	2.457（.007）	3.355	.983
7	60	-1.708	-1.049	-1.378	2.948（.002）	9.679	.995
8	70	-1.708	-0.939	-1.324	3.439（.000）	33.867	.994

分析和讨论，选择这批数据的原因是可以将本文方法的结果和已有研究中的结果进行比较。进一步基于分析结果有针对性地设计了模拟研究，进一步评价所提出方法的表现。

4 实证数据分析

实证数据来自某个基于计算机施测的职业认证测验，关于该数据集的详细介绍可见 Cizek 和 Wollack（2017）。我们分析是这个数据集里的 Form1，包含 170 道二级计分的项目，有 1636 名被试在 170 道项目上的作答数据和作答时间数据，经过施测机构广泛的调查，标定这个 Form1 里包含 64 道泄漏的项目，并通过多种统计方法和检测程序，将其中的 46 名被试判定为可疑的作弊者。近年来，该数据集受到广泛关注，许多研究者都对该数据集进行过分析和研究（Kasli et al., 2023; Lee, 2018; Sinharay, 2017a, 2017b, 2020）。与以往研究一样，将统计量的检测结果与数据中存在的标签进行比较就可以对不同统计量的表现进行相对比较。

根据数据集所标出的泄漏题目，将整个测验分成两部分，即正常题目集合和泄漏的题目集合。为了更直观地考察原数据集中“标记”的异常考生在测验中是否存在作答速度上的差异，对 Form1 的数据进行如下的分析：（1）在剔除包含缺失数据和“标记”异常的数据之后，将对数正态时间模型对数据拟合，得到各题目的项目参数。（2）在剔除包含缺失数据的考生后，针对数据集中“标记”出的 41 名异常考生，分别基于测验中所有的 170 题、未泄漏的 106 题和泄漏的 64 题进行速度参数估计，得到的速度参数如图 2 所示。（3）在剔除包含缺失数据的考生之后，分别采用 SLR、BF 和 PP，对测验中的所有 1624 考生进行分析，并把“标记”异常的考生进行速度参数估计，如图 3、图 4 和图 5 所示。

从图 2 可以看出，有 24 名考生被标记为异常的考生在泄漏题目上的速度大于其在正常题目上的速度，这在图 1 上表现为红色显示的速度大于绿色显示的速度。与此同时，有 17 名考生则没有体现这种现象，这表明这部分被标记的考生并没有显示出明显的速度差异。为了进一步探查这部分考生的作答

表 3 三种统计量对实证数据的分析结果

	检测出的考生编号
BF	**251**，324，**452**，**453**，465，**524**，525，**624**，679，700，792，993，1252
SLR	**251**，324，**453**，465，**524**，525，**624**，679，700，792，993
PP	**251**，324，**453**，**524**，525，**624**，679，792，993

注：带灰色背景加粗显示的数字表示该考生在原数据中也被标示出为“异常”，其中 SLR，BF 和 PP 采用的临界值分别为 SLR 的 p 值小于 .05，BF 值大于 1 和 PP 值大于 .95。

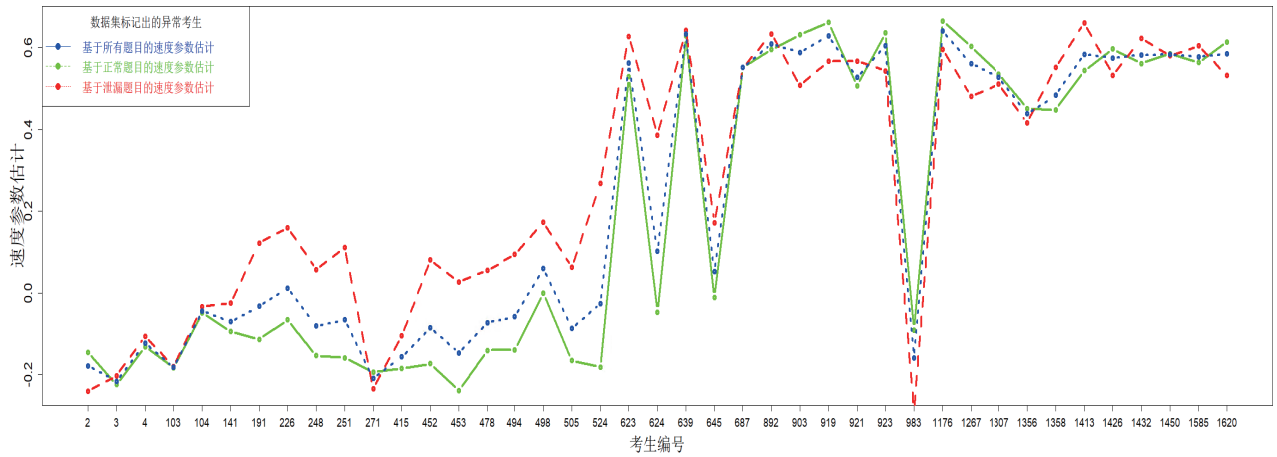

图 2 对原数据集标出的“异常考生”的速度参数估计

注：图中的三条线对应的是数据集中标出的“异常考生”的速度参数估计值，其中红色的线是基于数据集中标出的泄漏题目估计的结果，蓝色的线是基于数据集中所有题目估计的结果，绿色线是基于数据中未泄漏题目估计的结果。

行为，我们对他们得分数据进行具体分析，对他们的完整得分数据、正常题目的得分数据和泄漏题目上的得分数据分别进行能力参数估计。结果表明，这部分考生虽然没有显示出速度差异，但是他们显示出了能力差异，将他们在泄漏题目上的能力估计值$\hat{\theta}_1$减去他们在正常题目上的能力估计值$\hat{\theta}_2$，得到能力差异$\hat{\theta}_1 - \hat{\theta}_2$的平均值为 .818，能力差异的区间为 [.05, 2.09]，即基于他们在泄漏题目上的能力参数估计或多或少地都大于在正常题目上的能力参数估计。这里面的原因有可能是考生的作答策略或认知风格所导致的（王超，2018），有的考生更倾向于使用更保守的时间策略，或者考生可能有意识地对抗题目预知的影响。这个结果表明题目预知对不同考生的影响是不同的，也表明将题目预知对考生的影响按固定效应处理的方式是不恰当的。

进一步，针对 Form1 中的题目分类（未泄漏和泄漏的题目）和对应的作答时间数据，利用符号似然比 SLR 和前面构建的贝叶斯因子 BF 及后验概率 PP，对每位考生的作答时间数据进行检测，将结果与原数据集中的考生标注进行比较，具体结果如下表 3 所示。可以看出，相对于原数据集中标记出的“异常考生”，三种统计方法都更“保守”。BF，SLR 和 PP 分别只检测出了 13，11 和 9 个“异常考生”，并且这三个方法检测出的考生集合是包含关系，即 BF 检测出的结果包含了 SLR 和 PP 检测出的考生，SLR 检测出的结果包含了 PP 检测出的考生，即 PP 法在判定考生是否异常时在三种方法中相对最保守。

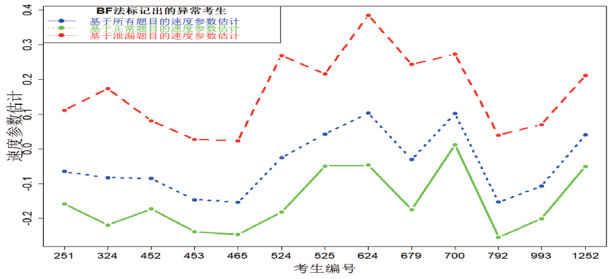

图 3 基于 BF 标出的“异常考生”的速度参数估计

并且，我们分析了这些考生的速度参数估计值，如图 3、图 4 和图 5 所示。可以看出，三种方法检测出的考生在泄漏题目上的作答速度大于其在正常题目上的作答速度，图中的三种颜色（红，绿和蓝，彩图见电子版）的线分别对应考生在泄漏题目上的速度、正常题目上的速度和整个测验的速度。在图 3、图 4 和图 5 上三条线保持了相同的顺序，即从上到下依次是红、蓝和绿，并且红色线距蓝色线的距离大于蓝色线距绿色线的距离，体现了考生在泄漏题目上存在题目预知对于考生作答速度的影响。

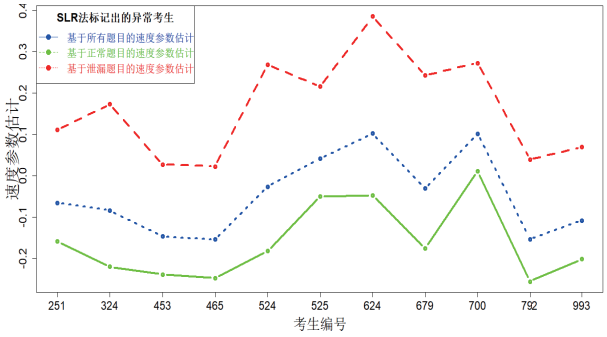

图 4 基于 SLR 标出的“异常考生”的速度参数估计

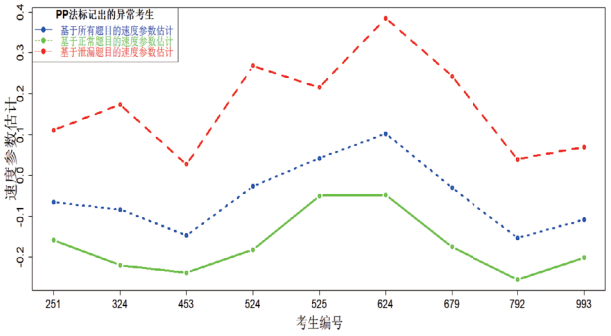

图 5 基于 PP 标出的“异常考生”的速度参数估计

从图 3，图 4 和图 5 所描述的结果来看，三种方法检测出的“异常考生”在泄漏题目上的作答速度大于其在正常题目上的作答速度，不同的是 PP 法检测出的人数最少，BF 法检测出的人数最多，但是 PP 没有检测出但是 BF 检测出的那些考生仍然存在速度差异，表明 BF，SLR 和 PP 法的检验力是从高到低。

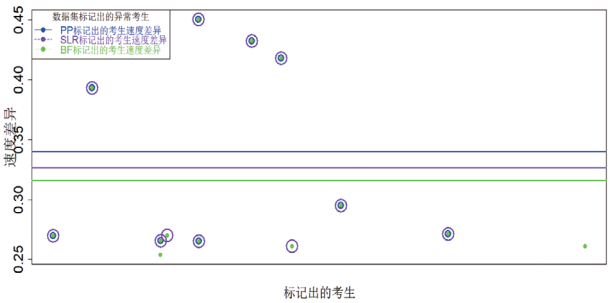

图 6 三种方法检测出的“异常考生”在泄漏题目和正常题目上的平均速度差异

图 6 描述了三种方法检测出的异常考生的作答速度差异和速度差异的均值，13 个绿色点是 BF 标记出的异常考生，其中有 11 紫色圆圈标出的是 SLR 标记出的考生，9 个蓝色圆圈标出的是 PP 标记出的考生。三条水平线分别对应不同方法标记出的考

生的速度差异均值。可以看出，BF，SLR 和 PP 方法检测出考生的速度差异顺序是从小到大，分别是：.316，.326，.34，即 BF 方法对速度差异最敏感，而 PP 法对速度差异相对最不敏感，这与前面表 3，图 3，图 4 和图 5 中得到的结果一致。

图 6 也进一步表明三种方法对于检测速度差异的敏感性存在差异，为了进一步评价 SLR，BF 和 PP 在不同测验条件下的表现，参考实证研究的分析结果，我们进一步开展了如下的模拟实验。

5 模拟实验

参考前面实证数据分析的结果，这里操纵的变量包括三个，分别是：被预知题目的比例、题目预知影响的考生人数比例和受题目预知影响的程度。

5.1 实验设计

测验长度参考 Sinharay 和 Johnson （2021）的设置，测验包括 100 个题目，项目参数的分布来自 Cizek 和 Wollack （2017），考生的速度参数从标准正态分布中抽取。三个操纵的变量：（1）被预知题目的比例，随机抽取固定比例（考虑 3 个水平，分别为 10%，20%，30%）的题目作为被预知的题目。（2）受题目预知影响的考生人数比例，也考虑 3 个水平，分别为 5%、10%、20%。考生的数量固定为 1000，因此有题目预知的考生数量分别为 50，100，200。（3）受题目预知影响的程度（即有题目预知的考生其速度参数所增加的量，这个值越大表明考生受题目预知影响的越严重），这里参考实证数据的分析，考察三个水平，分别是低 U（.20~.35）、中 U（.35~.50）和高 U（.50~1），表示受影响的考生在泄漏题目上的速度减去其在正常题目上的速度是从对应的均匀分布中抽取。这个设置是参考了前面实证数据分析的结果，相对于已有研究的速度差异设置为 1 或 2 （Wang et al., 2018; Zhu et al., 2023），这里采用的速度差异值相对较小，更接近实际应用中的情形。

综合来说，一共有种实验条件，如表 4 所示。速度差异水平和预知题目比例的不同组合，代表不同严重程度的题目预知。受影响的考生比例代表着题目预知在考生总体中的“流行程度”。对于每种实验条件重复生成 20 批数据，分别用三种方法对数据进行分析，并计算对应的评价指标。

5.2 评价指标

评价指标采用一类错误率和统计检验力，其中一类错误率表达的是方法将正常作答考生识别为有题目预知考生的比率，统计检验力是指方法成功识别有题目预知考生的比率。计算公式分别如下：

$$\text{统计检验力} = \frac{\text{正确识别受 preknowledge 影响的考生的数量}}{\text{受 preknowledge 影响的考生总数}}, \quad (10)$$

$$\text{I 类错误率} = \frac{\text{将正常考生错误标识为异常考生的数量}}{\text{正常作答考生总数}}, \quad (11)$$

其中，统计检验力越高表明选用的方法在识别有的考生的能力越强，I 类错误率越接近显著性水平表明方法对 I 类错误控制地越好。

5.3 数据生成过程

对于各考生在正常题目（没有预知的题目）上的作答时间是通过 van der Linden 的对数正态模型模拟，而对于考生在有预知的题目上的作答时间，我们参考前面的实证数据分析，考虑三种水平的速度差异，这样处理的目的是模拟考生在正常题目和预知题目上存在不同水平的速度差异。生成数据集后，分别计算 SLR 统计量和对应的 p 值、贝叶斯因子 BF 和考生的后验概率 PP，其中 SLR 根据 p 值进行判断，BF 的临界值采用 Kass 和 Raftery （1995）建议的标准，本研究中按考生的 BF 值大于 1 则判断考生存在题目预知，PP 的临界值采用 .95。

5.4 研究结果

SLR 的 p 值越小，SLR 或 PP 值越大都表明考生在测验和中存在速度差异的可能性越大。表 5 给出了 BF、SLR 和 PP 三种方法在不同条件下检测考

表 4 实验设计

因素	水平数（取值）
预知题目占测验总题目的比例	3 （.1, .2, .3）
受预知题目影响的考生比例	3 （.05, .1, .2）
速度受题目预知影响的程度	3 （U（.20-.35）, U（.35-.50）, U（.50-1））

生存在题目预知时的统计检验力和一类错误率。可以看出，当存在题目预知考生的速度差异水平较低“U（.20-.35）”，并且预知题目的比例也较低（.1）时，各方法的检验力相对都较低，不超过 .6，但是当速度差异水平达到较高“U（.5-1）”时，各方法的检验力可以达到较高的水平，最高可以达到 .99 以上。参考实证数据的结果，速度差异中等时最接近实证数据，这种情况下，BF 和 SLR 在最轻微的题目泄漏条件（泄漏题目比例 .1）下，检验力都能达到 80% 以上。随着预知题目比例的增加，三种方法的检验力会随之增加，因为相对较多的泄漏题目有利于对受影响考生的检验。本研究是基于题目参数已知条件下开展的，因此受影响的考生比例对于检验力影响较小。并且，从结果来看，三种方法在各条件下都能较好的控制一类错误率。综合来看，BF 方法在统计检验力上表现最好，在各条件下检验的准确率最高，SLR 方法的检验力相近但是略低于 BF，PP 法的检验力在三种方法是最低的，这与其对速度差异的不敏感有直接关系。

图 7 展示了题目预知考生的速度差异水平为 U（.20~.35）时，三种预知题目比例（.1, .2, .3）条件下，三种方法的统计检验力的比较。容易看出，随着预知题目比例的增加，各方法的检验力都有增加。预知题目比例从 .1 增加到 .2 的时候，各方法增加的幅度在 15% 左右，最大的增幅超过 20%，而预知题目比例从 .2 增加到 .3 的时候，增加的幅度不超过 10%，最小的增幅不到 3%。表明要想达到较高准确率的检测考生的题目预知，需要达到预知 20% 的题目的水平。

表 5 BF，SLR 和 PP 三种在不同条件下检测题目预知考生的表现

速度差异分布	预知题目的比例	受影响的考生比例	统计检验力			一类错误率		
			BF	SLR	PP	BF	SLR	PP
低 U（.20-.35）	.1	.05	.506	.502	.388	.050	.046	.040
		.1	.492	.485	.372	.048	.044	.039
		.2	.517	.493	.392	.051	.048	.039
	.2	.05	.682	.632	.590	.052	.037	.045
		.1	.714	.661	.589	.048	.033	.040
		.2	.734	.678	.609	.050	.038	.047
	.3	.05	.820	.760	.728	.049	.032	.043
		.1	.819	.754	.742	.054	.034	.047
		.2	.800	.739	.721	.047	.032	.048
中 U（.35-.50）	.1	.05	.830	.820	.636	.049	.045	.043
		.1	.811	.802	.597	.050	.046	.039
		.2	.803	.797	.598	.047	.045	.041
	.2	.05	.966	.950	.812	.049	.034	.042
		.1	.964	.958	.814	.051	.036	.043
		.2	.956	.936	.815	.049	.035	.045
	.3	.05	.990	.986	.868	.047	.029	.044
		.1	.986	.974	.885	.049	.033	.048
		.2	.987	.978	.875	.052	.032	.045
高 U（.50-1）	.1	.05	.806	.794	.606	.047	.044	.042
		.1	.794	.782	.596	.050	.048	.034
		.2	.817	.807	.616	.049	.045	.039
	.2	.05	.976	.952	.838	.051	.034	.043
		.1	.967	.945	.832	.050	.037	.042
		.2	.953	.935	.797	.046	.035	.044
	.3	.05	.992	.990	.890	.050	.031	.048
		.1	.990	.981	.880	.053	.033	.046
		.2	.984	.979	.865	.051	.031	.045

注：BF、SLR 和 PP 分别表示基于作答时间数据贝叶斯因子，符号似然比和后验概率统计量。

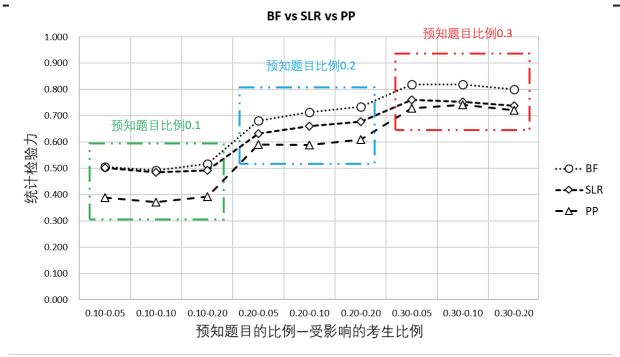

图 7　三种方法基于作答时间检测题目预知表现的比较

注：图中 X 轴上的数据标签中的两个数字分别表示预知题目的比例和受题目预知影响的考生比例，比如 .10–.05 表示预知题目的比例为 .10，受题目预知影响的考生比例为 .05。Y 轴表示检验题目预知的统计检验力。

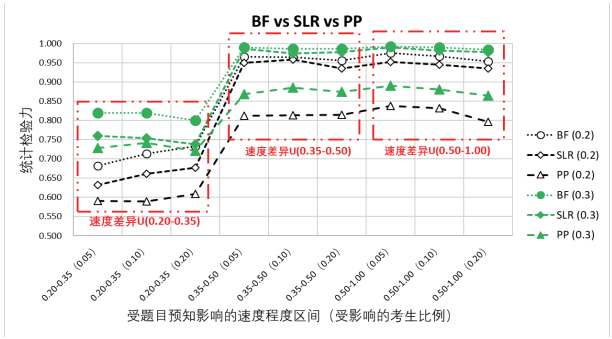

图 8　三种方法基于作答时间检测题目预知表现的比较

注：图中 X 轴上的数据标签中的数字分别表示受题目预知影响的速度程度区间和受题目预知影响的考生比例，比如 .20~.35（.05）中的 .20~.35 表示受题目预知影响的速度从均匀分布 U（.20~.35）中抽取，受题目预知影响的考生比例为 .05。Y 轴表示检验题目预知的统计检验力。

图 8 展示了两种预知题目比例（.2, .3）条件下，三种方法在不同的速度差异水平“低 U（.20, .35），中 U（.35, .50），高 U（.50, 1）”下的统计检验力的比较。在题目预知比例 .2 时，BF 方法在三种速度差异水平下的检验力均值分别为：.710，.962，.965；SLR 的均值为 .657，.948，.944；PP 的均值为 .596，.814 和 .822。容易看出，随着速度差异水平从 U（.20, .35）增加到 U（.35, .50），各方法的检验力都有较大的增加；而从 U（.35, .50）增加到 U（.50, 1）时，各方法的检验力基本保持稳定。表明各方法在速度差异达到中等水平时，就可以较准确地检测出有题目预知的考生。

6　讨论和进一步的研究方向

本文将贝叶斯因子和后验概率统计量进行拓展，构建了两个可以在作答时间数据下检测考生题目预知的统计量，并将它们与符号似然比检验统计量进行比较，利用它们来检测考生在泄漏题目和正常题目上作答速度的差异。在职业资格测验实证数据上的分析结果表明，当考生存在题目预知时，其在泄漏题目和正常题目上的作答速度存在差异，三个统计量（SLR, BF 和 PP）在检测时存在差异，其中 BF 对于速度差异最敏感，PP 对于速度差异最不敏感。需要注意的是原数据集中标记的部分考生并没有显示出速度差异，但是他们显示出了或多或少的能力差异，这表明题目预知对于不同考生所带来的影响是不同的，多数考生会表现出速度加快和正确作答概率升高，但是也有少数考生并不会表现出明显的速度变化，这也表明题目预知对不同考生的影响是不同的。在实证数据分析结果的基础上，有针对性地进一步设计了模拟实验，考察“题目预知带来不同程度的速度差异、题目预知在考生群体中的流行程度、预知题目占总测验题目的比例”三个因素，在不同测验条件下对它们的表现进行了综合比较。

模拟实验的结果进一步表明：（1）BF 对于速度差异最敏感，它的检验力略高于 SLR，PP 对于速度差异最不敏感；（2）三种方法都能较好地控制一类错误率；（3）在速度差异达到中等“U（.35, .50）”时，三种方法可以达到较高的检验准确率；（4）要想达到较准确的检测存在题目预知的考生，需要考生预知的题目达到 20% 或以上；（5）由于本研究是在已知题目参数条件下展开的研究，题目预知在考生群体中的流行程度对于检测结果影响较小。有两点需要说明，首先虽然本研究中所使用的实测数据为考生提供了作答行为标签（正常或异常），这并不是普遍现象，通常情况下的实测数据不会包含考生作答行为的标签，但是基于本文所提出方法在模拟和实证数据中的表现，表明它们可以在其它的实测数据中使用，尤其是大规模的测评数据，对考生的作答行为进行评价。其次在实际的异常作答行为检测中，测验实施者需要根据测验的性质和目的来选择合适的方法，比如在高利害的测验中，做决策时需要更谨慎，操作者可能会偏向于使用相对保守的检测方法，并且由于测验中往往不是单一的异常类型，很可能是多种不同异常类型的混合，因此需要采用多种方法来进行综合判断，尤其是针对目标异常类型进行专门检测。

虽然本研究对于检测测验中考生题目预知有较好的表现，但是仍存在一些局限，一些研究方向需要在未来的研究中考虑，主要表现在：（1）基于

时间数据的 BF 和 PP 的表现仍需要在更多的实证数据和模拟条件下探索，比如考察计算机自适应的测验场景下的表现。（2） 本研究中只是基于考生的时间数据展开的，将考生的得分数据和时间数据同时纳入 BF 和 PP，提高对考生在题目预知上的检验力，值得进一步探索和研究。（3）考察多维能力的测验在现实应用中十分常见，因此，将当前的 BF 和 PP 拓展到多维能力是未来需要开展的研究。（4）本研究中 SLR 和 PP 采用的是理论临界值，BF 采用的是经验临界值，这些临界值在实际的研究中是否合适以及探索更合适的临界值需要进一步研究。（5）在实际的应用中，有可能不完全掌握考生预知的题目信息，比如标定预知的 10 个题目中可能只有 8 道题被考生预知（Belov, 2016），因此，将预知题目信息的不确定性纳入考虑，是未来需要考虑的研究。

参考文献

胡佳琪 , 黄美薇 , 骆方 . (2020). 考试作弊甄别技术的研究进展：个体作弊的甄别 . *中国考试 , 11,* 32–36.

黄庆 , 王倩 , 闻家君 , 涂冬波 . (2024). 加速作答行为的混合项目反应树模型研究 . *江西师范大学学报 (自然科学版), 48(*5), 449–458.

梁润美 . (2022). *基于反应时间的被试异常行为检测的变点分析方法比较* (硕士学位论文), 东北师范大学，长春 .

刘玥 , 刘红云 . (2021). 心理与教育测验中异常作答处理的新技术 : 混合模型方法 . *心理科学进展 , 29*(9), 1696–1710.

刘玥 , 刘红云 , 游晓锋 , 杨建芹 . (2022). 用于处理不努力作答的标准化残差系列方法和混合多层模型法的比较 . *心理学报 , 54*(4), 411–425.

李亚玲 . (2021). *心理与教育测验中侦测表现下降新视角—基于 JS 散度的变点分析法* (硕士学位论文), 江西师范大学，南昌 .

骆方 , 王欣夷 , 徐永泽 , 封慰 . (2020). 考试作弊甄别技术的研究进展：团体作弊的甄别 . *中国考试 , 11*, 37–41.

秦春影 , 吴龙月 , 王爱平 . (2022). 计算机自适应测验中试题泄露的实时监控方法研究与应用 . *江西师范大学学报 (自然科学版), 46*(2), 118–125.

童昊 , 喻晓锋 , 秦春影 , 彭亚风 , 钟小缘 . (2022). 多级计分测验中基于残差统计量的被试拟合研究 . *心理学报 , 54(*9), 1122–1136.

王超 . (2018). *自适应测验中认知风格对作答时间的影响机制* (硕士学位论文), 山东师范大学，济南 .

王丹 , 刘红云 . (2023). 量表数据中不努力作答的识别和清理 . *心理学探新 , 43(*6), 558–566.

王雪 , 罗芬 , 蔡艳 , 涂冬波 . (2024). 迫选测验中后程随机作答的侦查：基于变点分析法 . *心理科学 , 47(*6), 1507–1518.

杨志明 , 徐庆树 . (2023). 基于项目作答反应时间的作弊甄别研究进展 . *心理学探新 , 43*(3), 278–288.

钟小缘 , 喻晓锋 , 苗莹 , 秦春影 , 彭亚风 , 童昊 . (2022). 基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数 . *心理学报 , 54(*10), 1277–1292.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement, 40*(2), 83–97.

Cheng, Y., & Shao, C. (2021). Application of change point analysis of response time data to detect test speededness. *Educational and Psychological Measurement, 82*(5), 10311062.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests.* Routledge.

Fox, J.–P., Koops, J., Feskens, R., & Beinhauer, L. (2020). Bayesian covariance structure modelling for measurement invariance testing. *Behaviormetrika, 47*(2), 385–410..

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4),540–553.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80(*2), 312–345.

Hong, M., Lin, L., & Cheng, Y. (2021). Asymptotically corrected person fit statistics for multidimensional constructs with simple structure and mixed item types. *Psychometrika, 86*(2), 464–488.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Kasli, M., Zopluoglu, C., & Toton, S. L. (2023). A deterministic gated lognormal response time model to identify examinees with item preknowledge. *Journal of Educational Measurement, 60*(1), 148–169.

Lee, S. Y. (2018). *A Mixture model approach to detect examinees with item preknowledge* (Doctorial dissertation). University of Wisconsin–Madison.

Lu, Y., & Sireci, S. G. 2007. Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(3), 29–37.

Marianti, S., Fox, J. P., Marianna, A., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426451.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Pan, Y. Q., & Wollack, J. A. (2021). An unsupervised–learning–based approach to compromised items detection. *Journal of Education Measurement, 58*(3), 413–433.

Robert, C. P. (2007). *The Bayesian choice.* Springer.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two–state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232.

Shao, C. (2016). Aberrant response detection using change–point analysis(Unpublished doctoral dissertation). University of Notre Dame.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika, 78*(3), 481–497.

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics, 41*(5), 521–549.

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46–68.

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403–421.

Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement, 44*(5), 376–392.

Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology, 73*(3), 397–419.

Sinharay, S., & Johnson, M. S. (2021). The use of the posterior probability in score differencing. *Journal of Educational and Behavioral Statistics, 46*(4), 403–429.

Stern, H. S. (2005). Model inference or model selection: Discussion of Klugkist, Laudy, and Hoijtink(2005). *Psychological Methods, 10*(4), 494–499.

van der Linden, W. J. (2006). A lognormal model for response times on test forms. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.

van der Linden, W. J. (2009). Conceptual issues in response–time modeling. *Journal of Educational Measurement, 46*(3), 247–272.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120–139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response–time patterns in adaptive testing. *Psychometrika, 73*(3), 365384.

van der Linden, W. J., van Krimpen–Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251–265.

van Krimpen–Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM–based person–fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*(2), 199–218.

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling, 53*(3), 334–358.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics, 43*(4), 469–501.

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement, 41*(4), 243–263.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer–based tests. *Applied Measurement in Education, 18*(2), 163183.

Wollack, J. A., & Schoenig, R. W. (2018). *The Sage encyclopedia of educational research, measurement, and evaluation*. Sage.

Zhu, H. Y., Jiao, H., Gao, W., & Meng, X. B. (2023). Bayesian change–point analysis approach to detecting aberrant test–taking behavior using response times. *Journal of Educational and Behavioral Statistics, 48*(4), 490–520.

Speed Difference Detection Based on Response Time Data

Xin Yunxi[1], *Qin Chunying*[1,2], *Dong Shenghong*[1], *Yu Xiaofeng*[1]

([1] School of Psychology, Jiangxi Normal University, Nanchang, 330022)

([2] School of Mathematics and Information Science , Nanchang Normal University, Nanchang, 330032)

Abstract Response time data is increasingly recognized for its potential to reveal the pace and conduct of examinees, offering valuable insights into educational and psychological assessments. Unusually rapid test completion may suggest irregular behavior, such as obtaining prior knowledge of certain test items ahead of his/her test. Current research indicates that the signal likelihood ratio (SLR) test outperforms other methods in maintaining type I error rates and enhancing statistical detection powers. This paper focuses on comparing the SLR with two novel test statistics designed to detect speed discrepancies.

Using response time data, we developed two Bayesian-inspired statistics to assess variations in test-taking speed. To gauge the efficacy of these statistics in detecting prior knowledge of test items, we initiate our analysis with a well-known data set from real-world scenarios. This data set has been previously scrutinized in studies aimed at identifying item preknowledge, allowing us to benchmark our findings against existing literature. Employing the signal likelihood ratio (SLR), the Bayesian factor (BF), and the posterior probabilistic (PP) approaches, we scrutinize each examinee's response time data. Based on the flagged items marked in the dataset, the entire test is divided into two parts: the collection of normal items and the collection of flagged items. To more intuitively examine whether the "marked" abnormal examinees in the original dataset have differences in response speed during the test, the data of Form1 is analyzed as follows: (1) After excluding data with missing information and "marked" abnormal data, the log-normal time model is fitted to the data to obtain the item parameters for each item. (2) After excluding examinees with missing data, for the 41 "marked" examinees in the dataset, speed parameter estimation is conducted based on all 170 questions in the test, 106 non-leaked questions, and 64 leaked questions. The speed parameters obtained indicate that the examinees identified by the three methods all have speed differences. (3) After excluding examinees with missing data, all 1624 examinees in the test are analyzed using SLR, BF, and PP, and speed parameter estimation is conducted for the "marked" abnormal examinees.

The outcomes are then juxtaposed with the marked "aberrant examinees" in the original data set. Interestingly, all three methods exhibit a more "conservative" stance compared to the original dataset's annotations. Specifically, BF, SLR, and PP identify 13, 11, and 9 examinees, respectively. Moreover, the detection sets from these methods are inclusive, with BF encompassing both SLR and PP detection, and SLR encompassing PP detection. This suggests that the PP method is the most stringent in flagging abnormal examinee behavior, while BF is comparatively more lenient.

Building on these findings, we designed a simulation study to further appraise the performance of the proposed methods. The results indicate that examinees with prior knowledge exhibit distinct response speeds on leaked versus normal items. The sensitivity of the three statistics—SLR, BF, and PP—varies, with BF being the most responsive to speed differences and PP the least. Targeted simulation experiments are conducted to assess the impact of varying degrees of speed differences due to item preknowledge, the prevalence of such preknowledge among examinees, and the proportion of known items within the total test under diverse conditions. A comprehensive comparison reveals that: (1) All three methods effectively control type I error rates; (2) A medium speed difference (U = .35 to .50) allows for high detection accuracy; (3) To accurately identify examinees with item preknowledge, they must have prior knowledge of at least 20% of the items; and (4) Given the study's parameters are known, the prevalence of preknowledge in the population is expected to have a minimal impact on test outcomes. The newly developed statistics demonstrate robust performance in detecting response speed differences during the examination process.

Key words response time, speed, posterior probability, difference detection, Bayesian factor